Coded Caching With Private Demands and Caches

Ali Gholami*, Kai Wan*, Hua Sun†, Mingyue Ji‡, Giuseppe Caire*
*Technische Universität Berlin, 10587 Berlin, Germany, {ali.gholami, kai.wan, caire}@tu-berlin.de
†University of North Texas, Denton, TX 76203, USA, hua.sun@unt.edu
‡University of Utah, Salt Lake City, UT 84112, USA, mingyue.ji@utah.edu

2022 IEEE International Symposium on Information Theory (ISIT) | 978-1-6654-2159-1/22/$31.00 ©2022 IEEE | DOI: 10.1109/ISIT50566.2022.9834846

***Abstract*—In the coded caching literature, the notion of privacy is considered only against demands. On the motivation that multi-round transmissions almost appear everywhere in real communication systems, this paper formulates the coded caching problem with private demands and caches. Only one existing private caching scheme, which is based on introducing virtual users, can preserve the privacy of demands and caches simultaneously, but at the cost of an extremely large subpacketization exponential in the product of the number of users (K) and files (N) in the system. In order to reduce the subpacketization while satisfying the privacy constraints, we propose a novel approach which constructs private coded caching schemes through private information retrieval (PIR). Based on this approach, we propose novel schemes with private demands and caches which have a subpacketization level in the order exponential with K instead of NK in the virtual user scheme. As a by-product, for the coded caching problem with private demands, a private coded caching scheme could be obtained from the proposed approach, which generally improves the memory-load tradeoff of the private coded caching scheme by Yan and Tuninetti.**

***Index Terms*—Coded caching, private demands and caches, private information retrieval**

I. Introduction

The seminal work of Maddah-Ali and Niesen on coded caching (MAN) [1], brought up a new method to leverage from the local memories available on end users' devices to reduce the network load significantly. Two phases, *placement* and *delivery*, are included in a coded caching scheme. Let the library have N files. In the placement phase, each of the K users in the system stores M files in size in its cache without knowledge of future demands. In the delivery phase, the demands of all the users are revealed to the server, and the server sends multiple multicast messages to users to ensure each user can recover its demand. The objective is to minimize the transmission load from the server. The literature on coded caching includes decentralized coded caching [2], online coded caching [3], coded caching with random demands [4], D2D coded caching [5], hierarchical coded caching [6], coded caching with nonuniform demands [7], etc.

The coded caching problem with private demands was considered in [8], [18]. Two coded caching schemes were proposed in [8], from which each user cannot obtain any information about other users' demands from the transmission of the server. The first method introduces $K(N-1)$ virtual users to the system so that each file is requested by exactly K users in total. In this way, from the viewpoint of each real user, the multicast messages will be symmetrical and thus the demands are hidden. To reduce the subpacketization of the virtual user scheme (which is exponential in NK), an MDS precoding was proposed to make the multicast messages symmetric. An alternative scheme based on the virtual user strategy was proposed in [9], which is also with the subpacketization exponential in NK. Another strategy to preserve the demand privacy was introduced in [10], by storing some keys at the users' caches which are unknown to the other users. The main advantage of the scheme in [10] is its subpacketization which is the same as the original MAN caching scheme (which is exponential in K). Other papers on demand privacy include [11]–[14], which we leave them to the readers.

In a PIR setting, there is a user who is willing to retrieve a message from multiple databases. The objective is that the databases should not gain any information about the user's demand from the query and answer on that database and at the same time to keep the communication load minimum. In [16], the capacity of PIR has been characterized. For other works on PIR, one can refer to [17], [20], and [21].

In this paper, we are focusing on the privacy of the users' demands and caches simultaneously. The motivation stems from the fact that, in practice, users make sequential demands in a multi-round fashion (e.g., in video streaming, each video chunk is a "demand"). So our goal is to make it possible for multi-round caching to remain private in demand, which only happens when cache privacy is also preserved. Till now, the privacy issue of users' caches has not been considered in the literature. Besides the novel problem formulation, our main contributions are as follows.

- We first show that the virtual user scheme in [9] is private in terms of users' demands and caches.
- To reduce the subpacketziation of the virtual user scheme in [9], we propose a novel approach which constructs coded caching schemes with private demands and caches through PIR; that is, given any PIR scheme satisfying some conditions, we can construct a private coded caching scheme. For the case that the number of files is no more than 4, we obtain coded caching schemes with private demands and caches with a significantly reduced subpacketization compared to the scheme in [9]. We also propose a scheme for general N, which leads to lower subpacketization than the scheme in [9].
- As a by-product, for the coded caching problem with private demands, using the optimal PIR scheme in [16] into our approach results in a coded caching scheme with only the demand privacy requirement, which generally improves the memory-load tradeoff of the scheme in [10].

II. System Model and Related Results

We consider a shared-link coded caching system with a server connected to K cache-aided users. The server has access to a library of N independent files, denoted by $W_1, W_2, \dots, W_N$, each of which has F bits. We denote the library by $W_{[N]}$. The caching system operates in two phases.

Placement Phase. Each user fills its cache without knowledge of later demands. The cached content of user $k \in [K]$ is

$$Z_k = \phi_k(W_1, \dots, W_N, \mathscr{M}_k), \tag{1}$$

where $\mathscr{M}_k$ represents the metadata/composition of the bits in Z_k. $\mathscr{M}_k$ is a random variable over $\mathscr{C}_k$, representing all types of cache placements of user k. The realization of $\mathscr{M}_k$ is given to the server and not to other users. The memory size constraint is expressed by

$$H(Z_k) \leq MF, \ \forall k \in [K]. \tag{2}$$

As in [8], we assume that the length of $\mathscr{M}_k$ is negligible compared to the file size.

Delivery Phase. During the delivery phase, each user requests one file W_{d_k}, where d_k is distributed uniformly i.i.d. over $[n]$. Given the demand vector $\mathbf{d} = (d_1, \dots, d_K)$, the server broadcasts to all users the message

$$X_{\mathbf{d}} = \psi(\mathbf{d}, W_1, \dots, W_N, \{\mathscr{M}_j : j \in [K]\}), \tag{3}$$

Note that we have

$$\begin{aligned} H(W_{[N]}, \mathscr{M}_1, \dots, \mathscr{M}_K, \mathbf{d}) = NF + H(\mathscr{M}_1, \dots, \mathscr{M}_K) \\ + \sum_{k \in [K]} H(d_k). \end{aligned} \tag{4}$$

Decoding. User $k \in [K]$ decodes its desired file W_{d_k} from $\big(d_k, Z_k, X_{\mathbf{d}}\big)$, i.e.,

$$H\big(W_{d_k} | d_k, Z_k, X_{\mathbf{d}}\big) = 0. \tag{5}$$

Privacy. We want to preserve the privacy of each user's demand against other users as in [8], i.e.,

$$I(\mathbf{d}; X_{\mathbf{d}} | d_k, Z_k) = 0, \ \forall k \in [K]. \tag{6}$$

In addition to (6), we want to preserve the privacy of the metadata of each user's cache against other users, i.e.,

$$I\big((\mathscr{M}_1, \dots, \mathscr{M}_K); X_{\mathbf{d}} | d_k, Z_k\big) = 0, \ \forall k \in [K]. \tag{7}$$

Objective. We say that the load R is achievable if there exist cache placement functions $\phi_k(.)$, encoding function $\psi(.)$, and decoding functions $\theta_k(.)$ such that

$$W_{d_k} = \theta_k(d_k, Z_k, X_{\mathbf{d}}), \forall k \in [K], \tag{8}$$

$$\text{where } H(X_{\mathbf{d}}) \leq RF. \tag{9}$$

Our objective is to find the minimum achievable load $R^\star$ for given system parameters M, N, K,[1] i.e.,

$$R^\star = \min_{\phi_k, \psi, \theta_k : k \in [K]} R. \tag{10}$$

A. Preliminary Result

Theorem 1 (Upper bound on load)**.** *The load defined in* (10) *for a private caching system is upper bounded by the load of the virtual user scheme in [9], i.e.*

$$R^\star \leq R^p(N, K, M) := \frac{\binom{NK}{KM+1} - \binom{NK-N}{KM+1}}{\binom{NK}{KM}}. \tag{11}$$

Proof. To prove this theorem, we need to show that the virtual user scheme in [9] satisfies also the cache privacy constraint in (7), since the demand privacy constraint was already proved to be satisfied in [9]. Note that in this virtual user scheme, user k acts as user $\big((k-1)N + S_k\big)$ in the (N, NK, M) non-private scheme in [15]. Hence, since the variable $(k-1)N + S_k$ reveals the cache content of user k, it is clear that the metadata of cache defined in our scheme for the k-th user is $\mathscr{M}_k = S_k$ in which $S_k \sim \text{Unif}[N]$. So the left handside of (7) will be equal to $I\big((S_1, \dots, S_K); X_{\mathbf{d}} | d_k, S_k, Z_k\big)$ which equals 0 since the variables $S_1, \dots, S_K$ are independent of $X_{\mathbf{d}}$. □

B. Brief Review on the PIR Model

In the following, we briefly review the setting of the PIR problem in [16]. A user is willing to retrieve a file among N messages with B bits each (denoted by $W'_1, \dots, W'_N$) from L servers, each of which has access to all the N files. Assuming the desired file is W'_n, where n is uniformly distributed on $[N]$,[2] the user sends a query $Q_\ell^{[n]} \in \mathcal{Q}_\ell$ where $\ell \in [L]$ to servers ℓ. According to the received query, server ℓ sends back the answer $A_\ell^{[n]}$ as a function of the query and the messages $W'_1, \dots, W'_N$, to the user; i.e.,

$$A_\ell^{[n]} = \gamma_\ell(Q_\ell^{[n]}, W'_1, \dots, W'_N), \tag{12}$$

where γ_ℓ represents the encoding function of server ℓ. Note that the query $Q_\ell^{[n]}$ and the answer $A_\ell^{[n]}$ are not given to the servers in $[L] \setminus \{\ell\}$. In the decoding phase, the user should recover W'_n from the received answers, i.e.,

$$H\big(W'_n | A_1^{[n]}, \dots, A_L^{[n]}, Q_1^{[n]}, \dots, Q_L^{[n]}\big) = 0. \tag{13}$$

For the sake of privacy, each server cannot get any information about which message is demanded by the user, i.e.,

$$I(n; Q_\ell^{[n]} | W'_1, \dots, W'_N) = 0, \ \forall \ell \in [L]. \tag{14}$$

Note that from (14), we have $H(A_\ell^{[1]}) = \dots = H(A_\ell^{[N]}) := H(A_\ell)$, where $H(A_\ell)/L$ is defined as the transmission load of server ℓ. The transmission rate of a PIR scheme is defined as $\frac{B}{\sum_{\ell \in [L]} H(A_\ell)}$. Finally, we introduce an additional constraint on PIR which defines a new family of PIR schemes.

Definition 1. *For a* 2*-server PIR scheme, if the demand is uniformly distributed over* $[N]$ *and*

$$I(Q_1^{[n]}; Q_2^{[n]} | W'_1, \dots, W'_N) = 0, \ \forall n \in [N], \tag{15}$$

then the PIR scheme is a 2*-server PIR scheme with uniform demand and independent queries.*

[1] Note that the transmitted loads for different demands should be the same, by the constraint of private demands.

[2] Note that in [16], uniformity of the distributuion is not necessary, but it is needed in the caching problem.

III. Proposed Approach

Theorem 2. *Given any 2-server N-message PIR scheme with uniform demand and independent queries whose transmission rate is r, there exists an (N,K) coded caching scheme (N files and K users) with private demands and caches whose achieved memory-load tradeoff is the lower convex envelope of $(0,N)$,*

$$(M,R)=\left(\frac{Nt}{K}+\left(1-\frac{t}{K}\right)\frac{1}{2r},\frac{1}{2r}\frac{K-t}{t+1}\right),\forall t\in[0:K-1], \tag{16}$$

and $(N,0)$. Assume the needed subpacketization of the given PIR scheme is F', then the needed subpacketization for each point in (16) *with $t\in[0:K-1]$ is $\binom{K}{t}F'$.*

Intuitively, inspired by the coded caching scheme with private demands in [10], our proposed approach for Theorem 2 is based on the MAN coded caching scheme, and lets each user additionally cache some keys which are constructed by the answer of server 1 of the given PIR scheme. In the delivery phase, we transmit the messages on coded subfiles, which are encoded by the answer of server 2 of the given PIR scheme. It will be explained later that, the scheme in [10] can be seen as a special construction through our proposed approach.

Proof. We are given a 2-server PIR scheme in which the queries sent to the two servers are conditionally independent variables according to (15). Assume that the set of possible queries sent to the two servers are $\mathcal{Q}_1$ and $\mathcal{Q}_2$, respectively. Without loss of generality, we assume that the transmission loads of the two servers are the same; otherwise, a time-sharing step could be used.

Placement. For each $t\in[0:K-1]$, same as the MAN scheme, each file is splitted to $\binom{K}{t}$ nonoverlapping subfiles of equal length, i.e. for file $W_n, n\in[N]$,

$$W_n=(W_{n,\tau}:\tau\subset[K],|\tau|=t),$$

where each subfile contains $F/\binom{K}{t}$ bits. For every $n\in[N]$, $W_{n,\tau}$ is placed in the cache of user $k\in[K]$ if $k\in\tau$. Besides, for each user $k\in[K]$, we use the 2-server PIR scheme independently to randomly select a query $Q_{1,k}\in\mathcal{Q}_1$. Then we let each user store $\phi_1(Q_{1,k},W_{1,\tau},\ldots,W_{N,\tau})$ in its cache for each $\tau\subseteq[K]\backslash\{k\}$ where $|\tau|=t$. Since the PIR scheme is symmetric on the transmission rates of the two servers, it can be seen that $\phi_1(Q_{1,k},W_{1,\tau},\ldots,W_{N,\tau})$ contains $\frac{F}{2r\binom{K}{t}}$ bits. So in total user k caches

$$Z_k=\{W_{n,\tau}:n\in[N],\tau\subset[K],|\tau|=t,k\in\tau\}$$
$$\bigcup\{\phi_1(Q_{1,k},W_{1,\tau},\ldots,W_{N,\tau}):\tau\subseteq[K]\setminus\{k\},|\tau|=t\},$$

totally containing $\frac{N\binom{K-1}{t-1}}{\binom{K}{t}}F+\frac{\binom{K-1}{t}}{2r\binom{K}{t}}F=\left(\frac{Nt}{K}+\frac{K-t}{2rK}\right)F=MF$, satisfying the memory size constraint.

Delivery. Recall that in the delivery phase of the MAN coded caching scheme for (N,K,t), for each subset $\mathcal{S}\subset[K]$ of cardinality $|\mathcal{S}|=t+1$, the server transmits $\oplus_{s\in\mathcal{S}}W_{d_s,\mathcal{S}\backslash\{s\}}$, where $\oplus$ stands for bitwise XOR.

In the delivery phase, for each subset $\mathcal{S}\subseteq[K]$ where $|\mathcal{S}|=t+1$, the server transmits

$$Y_{\mathcal{S}}=\Sigma_{k\in\mathcal{S}}\phi_2\big(Q_{2,k}^{(d_k)},W_{1,\mathcal{S}\backslash\{k\}},\ldots,W_{N,\mathcal{S}\backslash\{k\}}\big). \tag{17}$$

Note that $Q_{2,k}^{(d_k)}$ is the query sent to server 2 in the PIR scheme, where the query $(Q_{1,k},Q_{2,k}^{(d_k)})$ corresponds to the $(d_k)^{\text{th}}$ message in the PIR problem. For each user $k\in\mathcal{S}$, since it can compute $\phi_2\big(Q_{2,k_1}^{(d_{k_1})},W_{1,\mathcal{S}\backslash\{k_1\}},\ldots,W_{N,\mathcal{S}\backslash\{k_1\}}\big)$ for each $k_1\in\mathcal{S}\setminus\{k\}$, it can recover $\phi_2\big(Q_{2,k}^{(d_k)},W_{1,\mathcal{S}\backslash\{k\}},\ldots,W_{N,\mathcal{S}\backslash\{k\}}\big)$. Then additionally with the cache content $\phi_1(Q_{1,k},W_{1,\mathcal{S}\backslash\{k\}},\ldots,W_{N,\mathcal{S}\backslash\{k\}})$, user k can recover $W_{d_k,\mathcal{S}\backslash\{k\}}$. Hence, by considering all subsets $\mathcal{S}\subseteq[K]$ where $|\mathcal{S}|=t+1$, each user can recover its desired file. In the delivery phase, the server totally transmits $\frac{\binom{K}{t+1}}{2r\binom{K}{t}}F=\frac{K-t}{2r(t+1)}F$, coinciding with (16).

The demand privacy is directly satisfied from the PIR scheme. More precisely, for each $k\in[K]$ we have

$$\begin{aligned}&I(\mathbf{d};(Y_{\mathcal{S}}:\mathcal{S}\subseteq[K],|\mathcal{S}|=t+1)|d_k,Z_k)\\&\leq I(\mathbf{d};(Y_{\mathcal{S}}:\mathcal{S}\subseteq[K],|\mathcal{S}|=t+1)|d_k,Z_k,W_{[N]}) &(18a)\\&\leq I(\mathbf{d};Q_{2,1}^{(d_1)},\ldots,Q_{2,K}^{(d_K)}|d_k,Z_k,W_{[N]}) &(18b)\\&=\sum_{k_1\in[K]\backslash\{k\}}I(d_{k_1};Q_{2,k_1}^{(d_{k_1})}|W_{[N]})=0 &(18c)\end{aligned}$$

where (18a) comes from (4), (18b) comes from that $(Y_{\mathcal{S}}:\mathcal{S}\subseteq[K],|\mathcal{S}|=t+1)$ is a function of $(Q_{2,1}^{(d_1)},\ldots,Q_{2,K}^{(d_K)})$ and $W_{[N]}$, (18c) comes from that each $(d_i,Q_{1,i},Q_{2,i}^{(d_i)})$ where $i\in[K]$ is independent of $(d_j,Q_{1,j},Q_{2,j}^{(d_j)})$ where $j\in[K]\backslash\{i\}$ given $W_{[N]}$ by our construction, and (18c) comes from (14).

Similarly, for the cache privacy constraint in (7), for each $k\in[K]$ we have

$$\begin{aligned}&I\big((\mathscr{M}_1,\ldots,\mathscr{M}_K);(Y_{\mathcal{S}}:\mathcal{S}\subseteq[K],|\mathcal{S}|=t+1)|d_k,Z_k\big)\\&\leq I\big((\mathscr{M}_1,\ldots,\mathscr{M}_K);(Y_{\mathcal{S}}:\mathcal{S}\subseteq[K],|\mathcal{S}|=t+1)\\&|d_k,Z_k,W_{[N]}\big) &(19a)\\&\leq I\big(Q_{1,1},\ldots,Q_{1,K};Q_{2,1}^{(d_1)},\ldots,Q_{2,K}^{(d_K)}|d_k,Z_k,W_{[N]}\big)\\& &(19b)\\&=\sum_{k_1\in[K]\backslash\{k\}}I\big(Q_{1,k_1};Q_{2,k_1}^{(d_{k_1})}|W_{[N]}\big)=0 &(19c)\end{aligned}$$

where again (19a) comes from (4), (19b) comes from that $\mathscr{M}_k=Q_{1,k},k\in[K]$ and that $(Y_{\mathcal{S}}:\mathcal{S}\subseteq[K],|\mathcal{S}|=t+1)$ is a function of $(Q_{2,1}^{(d_1)},\ldots,Q_{2,K}^{(d_K)})$ and $W_{[N]}$, and (19c) comes from that in the given PIR scheme each server has no information about the query of the other server.

□

Remark 1 (Subpacketization of the caching scheme for Theorem 2)**.** *If the needed subpacketization of the given PIR scheme for Theorem 2 is F', then the resulting coded caching scheme for Theorem 2 is $\binom{K}{t}F'$, where $t=\frac{KM}{N}\in[0:K]$.*

As a by-product of Theorem 2, given any 2-server PIR scheme for the original PIR problem as described in Section II-B, we can use the proposed approach to obtain a coded caching scheme with private demands. For a demand and cache private caching scheme, (15) should hold additionally.

Theorem 3. *Given any* 2*-server PIR scheme with* N *files and transmission rate* r*, there exists an* (N,K) *coded caching scheme with private demands whose achieved memory-load tradeoff is the lower convex envelope of* $(0,N)$*,* $(N,0)$*, and the points in* (16)*.*

Note that the PIR scheme in [17] achieves the optimal PIR rate, which is $r^\star := \frac{1}{1+1/2+(1/2)^2+\cdots+(1/2)^{N-1}}$, and needs the subpacketization equal to 1. Hence, by using the PIR scheme in [17] into Theorem 3, we have the following scheme.

Corollary 1. *For the* (N,K) *coded caching problem with private demands in [8], there exists a private-demand caching scheme whose achieved memory-load tradeoff is the lower convex envelope of* $(0,N)$*,*

$$(M,R) = \left(\frac{Nt}{K} + \left(1-\frac{t}{K}\right)\frac{1}{2r^\star}, \frac{1}{2r^\star}\frac{K-t}{t+1}\right), \forall t \in [0:K-1], \tag{20}$$

and $(N,0)$*. The needed subpacketization for each point in* (20) *with* $t \in [0:K-1]$ *is* $\binom{K}{t}$*.*

In addition, the coded caching scheme with private demands in [10] can be seen as a special case of Theorem 3. More precisely, by using the PIR scheme in [19] with transmission rate $1/2$ into Theorem 3, we obtain the scheme in [10]. Compared to the scheme in [10], since $r^\star > 1/2$, the proposed scheme in Corollary 1 has a strictly better performance on the memory-load tradeoff. In addition, the needed subpacketization of the two schemes are the same, equal to $\binom{K}{t}$ for each $t \in [0:K-1]$. A simple explicit example illustrating our construction in Theorem 2 can be found in [22].

IV. 2-Server PIR Schemes for Theorem 2

From our proposed approach in Theorem 2, to design coded caching schemes with private demands and caches, we only need to propose 2-server PIR schemes satisfying the query independence condition in (15); which we do in this section for different N. All the proposed 2-server PIR schemes have subpacketization level of $F' = 1$, which result in coded caching schemes with subpacketization of $\binom{K}{t}F' = \binom{K}{t}$ that is exponential in K, while the subpacketization of the virtual user scheme in [9] is exponential in NK. For $N = 2$, refer to [22], where we use the PIR scheme in Section III.A of [17].

A. $N = 3$

Let the file library be (A,B,C) with uniform demand distribution, and generate a random key T that is uniformly i.i.d. over the key set $\{0,1,2\}$. The proposed PIR scheme for different values of T and demands n is depicted in Table I.

Assume that $Q_i(Y)$ represent the query sent to server i to ask for the answer Y. Note that the queries in the proposed scheme are independent of file realization, so we can remove the terms in the condition from the constraints in (14) and (15).

	Server 1	Server 2		
		$n=A$	$n=B$	$n=C$
$T=0$	$A+B$	B	A	C
$T=1$	$A+C$	C	B	A
$T=2$	$B+C$	A	C	B

TABLE I: Proposed PIR scheme for $N = 3$.

The query to server 1 is clearly independent of the demand. For the query sent to server 2 denoted by Q_2, we have

$$\begin{aligned}\Pr(Q_2^{[n]} = Q_2(A)) &= \Pr(T=0, n=B) + \Pr(T=1, n=C)\\ &\quad + \Pr(T=2, n=A) = 1/3.\end{aligned}$$

In addition, we can compute $\Pr(Q_2^{[n]} = Q_2(A)|n = A) = \Pr(T = 2|n = A) = 1/3$. Hence, we have $\Pr(Q_2^{[n]} = Q_2(A)) = \Pr(Q_2^{[n]} = Q_2(A)|n = A)$. Similarly, we can prove that $P(Q_2^{[n]}) = P(Q_2^{[n]}|n)$; thus the query to server 2 is independent of the demand. Hence, (14) holds. For the constraint in (15), we have

$$\begin{aligned}&\Pr(Q_2^{[n]} = Q_2(A)|Q_1^{[n]} = Q_1(A+B))\\ &= \Pr(Q_2^{[n]} = Q_2(A)|T = 0) = 1/3 = \Pr(Q_2^{[n]} = Q_2(A)).\end{aligned}$$

Similarly, we can prove $P(Q_2^{[n]}|Q_1^{[n]}) = P(Q_2^{[n]})$; thus (15) holds. **The achieved transmission rate of this PIR scheme is $r = \frac{1}{2}$ and the subpacketization is $F' = 1$.**

For the example of the coded caching with private demands and caches with parameters $N = 3, K = 2, M = 2$, using this PIR scheme in Theorem 2 with $N = 3$ and $t = 1$, we get the achieved load of $\frac{1}{2}$ and subpacketization level of 2. In this example, the achieved load by the virtual user scheme in [9] is $\frac{2}{5}$ and the needed subpacketization level is 15.

B. $N = 4$

Let the file library be (A,B,C,D) with uniform demand distribution. Generate a random key T that is uniformly i.i.d. over the key set $\{0,1,2,3\}$. The proposed PIR scheme for different values of T and demands n is depicted in Table II.

The query to server 1 is clearly independent of the demand. For the query sent to server 2 denoted by Q_2, we have

$$\begin{aligned}&\Pr\left(Q_2^{[n]} = Q_2(-A+B+C+D)\right)\\ &= \Pr(T=0, n=A) + \Pr(T=1, n=B)\\ &+ \Pr(T=2, n=C) + \Pr(T=2, n=D) = 1/4.\end{aligned}$$

In addition, we can compute $\Pr\left(Q_2^{[n]} = Q_2(-A+B+C+D)|n = A\right) = \Pr(T = 0|n = A) = 1/4$. Hence, we have $\Pr\left(Q_2^{[n]} = Q_2(-A+B+C+D)\right) = \Pr\left(Q_2^{[n]} = Q_2(-A+B+C+D)|n = A\right)$. Similarly, we can prove that $P(Q_2^{[n]}) = P(Q_2^{[n]}|n)$; thus the query to server 2 is independent of the demand. Hence, (14) holds. For the constraint in (15), we have

$$\begin{aligned}\Pr\big(Q_2^{[n]} =& Q_2(-A+B+C+D)\\ &|Q_1^{[n]} = Q_1(A+B+C+D)\big)\\ &= \Pr(Q_2^{[n]} = Q_2(-A+B+C+D)|T=0)\\ &= 1/4 = \Pr(Q_2^{[n]} = Q_2(-A+B+C+D)).\end{aligned}$$

	Server 1	Server 2			
		$n = A$	$n = B$	$n = C$	$n = D$
$T = 0$	$A + B + C + D$	$-A + B + C + D$	$A - B + C + D$	$A + B - C + D$	$A + B + C - D$
$T = 1$	$-A - B + C + D$	$A - B + C + D$	$-A + B + C + D$	$A + B + C - D$	$A + B - C + D$
$T = 2$	$-A + B - C + D$	$A + B - C + D$	$A + B + C - D$	$-A + B + C + D$	$A - B + C + D$
$T = 3$	$-A + B + C - D$	$A + B + C - D$	$A + B - C + D$	$A - B + C + D$	$-A + B + C + D$

TABLE II: Proposed PIR scheme for $N = 4$.

	Server 1	Server 2				
		$n = A$	$n = B$	$n = C$	$n = D$	$n = E$
$T = 0$	$B + E$ $C + D$	A	E	D	C	B
$T = 1$	$A + B$ $C + E$	B	A	E	D	C
$T = 2$	$A + C$ $D + E$	C	B	A	E	D
$T = 3$	$A + D$ $B + C$	D	C	B	A	E
$T = 4$	$A + E$ $B + D$	E	D	C	B	A

TABLE III: Proposed PIR scheme for $N = 5$.

Similarly, we can prove $P(Q_2^{[n]}|Q_1^{[n]}) = P(Q_2^{[n]})$; thus (15) holds. **The achieved transmission rate of this PIR scheme is $r = \frac{1}{2}$ and the subpacketization is $F' = 1$.**

For the example of the coded caching with private demands and caches with parameters $N = 4, K = 2, M = \frac{5}{2}$, using this PIR scheme in Theorem 2 with $N = 4$ and $t = 1$, we get the achieved load of $\frac{1}{2}$ and subpacketization level of 2. In this example, the achieved load by the virtual user scheme in [9] is $\frac{1}{2}$ and the needed subpacketization level is 56.

C. General N

We will first illustrate our scheme with the example of $N = 5$. Let the file library be (A, B, C, D, E) with uniform demand distribution. Generate a random key T that is uniformly i.i.d. over the key set $\{0, 1, 2, 3, 4\}$. The proposed PIR scheme for different values of T and demands n is depicted in Table III.

For the general case of N, key T is uniformly distributed on $\{0, 1, \dots, N-1\}$. The answer of server 1 is comprised of multiple 2-sums of the files in library, i.e. $\{W_i + W_j, 1 \leq i < j \leq N\}$. The total number of 2-sums is $\binom{N}{2}$ and they are distributed among the N answers. So on average every answer of server 1 contains $\frac{N-1}{2}$ 2-sums. The set of answers of server 2 include 1-sums, i.e. $\{W_1, \dots, W_N\}$. To design the answer of server 1 and 2 for each T and n, we do reverse engineering as following. Consider in the first column for server 2 in the PIR table that corresponds to $n = 1$ as in Table III, we have the vector $(W_1; W_2; \dots; W_N)$ as the N answers for the N values of T. So for instance, in case of $T = 0$ and $n = 1$, server 2 sends W_1, for $T = 1$ and $n = 1$ it sends W_2, etc. For the next column that corresponds to $n = 2$, we shift the answer vector one unit downwards, i.e. $(W_N; W_1; \dots; W_{N-1})$. We continue this procedure and fill out all the columns of server 2. Now based on the answer of server 2 and the demand of the column n, we can design what is needed to be stored as the answer of server 1 for each T. For example, for $T = 0$ and $n = 1$, server 2 sends W_1 which is the demand, so nothing needs to be stored in this case. For $T = 0$ and $n = 2$, server 2 sends W_N but the user demands W_2. So, the 2-sum $W_2 + W_N$ should be in the answer of server 1 for $T = 0$. Continuing this process for each T and n, we fill out the answers of server 1 for different values of T.

The query to server 1 is clearly independent of the demand. For the query sent to server 2 denoted by Q_2, we have

$$\Pr(Q_2^{[n]} = Q_2(W_1)) = \Pr(T = 0, n = 1) + \Pr(T = 1, n = 2) + \cdots + \Pr(T = N-1, n = N) = 1/N.$$

In addition, we have $\Pr(Q_2^{[n]} = Q_2(W_1)|n = 1) = \Pr(T = 0|n = 1) = 1/N$. Hence, we have $\Pr(Q_2^{[n]} = Q_2(W_1)) = \Pr(Q_2^{[n]} = Q_2(W_1)|n = 1)$. Similarly, we can prove that $P(Q_2^{[n]}) = P(Q_2^{[n]}|n)$; thus the query to server 2 is independent of the demand. Hence, (14) holds. For the constraint in (15), for any $y \in \mathcal{Q}_1$ we have

$$\begin{aligned} &\Pr\left(Q_2^{[n]} = Q_2(W_1)|Q_1^{[n]} = Q_1(y)\right) \\ &= \Pr(Q_2^{[n]} = Q_2(W_1)|\text{some unique } T) = 1/N \\ &= \Pr(Q_2^{[n]} = Q_2(W_1)). \end{aligned}$$

Similarly, we can prove $P(Q_2^{[n]}|Q_1^{[n]}) = P(Q_2^{[n]})$; thus (15) holds. **The achieved transmission rate of this PIR scheme is $r = \frac{2}{N+1}$ and the subpacketization is $F' = 1$.**

Remark 2. *The resulting private coded caching scheme from the proposed PIR scheme in Section IV-C has a significant advantage on the subpacketization (with the order $2^{K\mathcal{H}(M/N)}$, where $\mathcal{H}(p)$ represents binary entropy function), compared to the private caching scheme in Theorem 1 (with the order $2^{NK\mathcal{H}(M/N)}$). However, the achieved load of the proposed caching scheme is linear with N. It is one of the on-going works to design schemes which achieves more flexible tradeoff between subpacketization and load.*

Acknowledgement The work of A. Gholami, K. Wan and G. Caire was partially funded by the European Research Council under the ERC Advanced Grant N. 789190, CARENET. The work of H. Sun was supported in part by NSF Awards 2007108 and 2045656. The work of M. Ji was supported in part by NSF Awards 1817154 and 1824558.

References

[1] M. A. Maddah-Ali and U. Niesen, ”Fundamental Limits of Caching,” IEEE Transactions on Information Theory, vol. 60, no. 5, pp. 2856-2867, May 2014.

[2] M. A. Maddah-Ali and U. Niesen, ”Decentralized Coded Caching Attains Order-Optimal Memory-Rate Tradeoff,” IEEE/ACM Transactions on Networking, vol. 23, no. 4, pp. 1029-1040, Aug. 2015.

[3] R. Pedarsani, M. A. Maddah-Ali and U. Niesen, ”Online Coded Caching,” IEEE/ACM Transactions on Networking, vol. 24, no. 2, pp. 836-845, Apr. 2016.

[4] M. Ji, A. M. Tulino, J. Llorca and G. Caire, ”Order-Optimal Rate of Caching and Coded Multicasting With Random Demands,” in IEEE Transactions on Information Theory, vol. 63, no. 6, pp. 3923-3949, June 2017, doi: 10.1109/TIT.2017.2695611.

[5] M. Ji, G. Caire and A. F. Molisch, ”Fundamental Limits of Caching in Wireless D2D Networks,” in IEEE Transactions on Information Theory, vol. 62, no. 2, pp. 849-869, Feb. 2016, doi: 10.1109/TIT.2015.2504556.

[6] N. Karamchandani, U. Niesen, M. A. Maddah-Ali and S. N. Diggavi, ”Hierarchical Coded Caching,” IEEE Transactions on Information Theory, vol. 62, no. 6, pp. 3212-3229, Jun. 2016.

[7] U. Niesen and M. A. Maddah-Ali, ”Coded Caching With Nonuniform Demands,” IEEE Transactions on Information Theory, vol. 63, no. 2, pp. 1146-1158, Feb. 2017.

[8] K. Wan and G. Caire, ”On Coded Caching With Private Demands,” IEEE Transactions on Information Theory, vol. 67, no. 1, pp. 358-372, Jan. 2021.

[9] Kamath, Sneha. ”Demand private coded caching.” arXiv preprint arXiv:1909.03324 (2019).

[10] Q. Yan and D. Tuninetti, ”Fundamental Limits of Caching for Demand Privacy Against Colluding Users,” IEEE Journal on Selected Areas in Information Theory, vol. 2, no. 1, pp. 192-207, Mar. 2021.

[11] Gurjarpadhye, Chinmay, et al. ”Fundamental limits of demand-private coded caching.” arXiv preprint arXiv:2101.07127 (2021).

[12] Aravind, V. R., Pradeep Kiran Sarvepalli, and Andrew Thangaraj. ”Coded Caching with Demand Privacy: Constructions for Lower Subpacketization and Generalizations.” arXiv preprint arXiv:2007.07475 (2020).

[13] S. Kamath, J. Ravi and B. K. Dey, ”Demand-Private Coded Caching and the Exact Trade-off for N=K=2,” National Conference on Communications (NCC), 2020, pp. 1-6.

[14] V. R. Aravind, P. K. Sarvepalli and A. Thangaraj, ”Subpacketization in Coded Caching with Demand Privacy,” National Conference on Communications (NCC), 2020, pp. 1-6.

[15] Q. Yu, M. A. Maddah-Ali and A. S. Avestimehr, ”The Exact Rate-Memory Tradeoff for Caching With Uncoded Prefetching,” IEEE Transactions on Information Theory, vol. 64, no. 2, pp. 1281-1296, Feb. 2018.

[16] H. Sun and S. A. Jafar, ”The Capacity of Private Information Retrieval,” IEEE Transactions on Information Theory, vol. 63, no. 7, pp. 4075-4088, Jul. 2017.

[17] C. Tian, H. Sun and J. Chen, ”Capacity-Achieving Private Information Retrieval Codes With Optimal Message Size and Upload Cost,” IEEE Transactions on Information Theory, vol. 65, no. 11, pp. 7613-7627, Nov. 2019.

[18] F. Engelmann and P. Elia, ”A content-delivery protocol, exploiting the privacy benefits of coded caching,” 15th Intern. Symp. on Modeling and Optimization in Mobile, Ad Hoc, and Wireless Networks (WiOpt), May 2017.

[19] N. B. Shah, K. V. Rashmi and K. Ramchandran, ”One extra bit of download ensures perfectly private information retrieval,” IEEE International Symposium on Information Theory (ISIT), 2014, pp. 856-860.

[20] Y. Zhou, Q. Wang, H. Sun and S. Fu, ”The Minimum Upload Cost of Symmetric Private Information Retrieval,” IEEE International Symposium on Information Theory (ISIT), 2020, pp. 1030-1034.

[21] K. Banawan and S. Ulukus, ”Multi-Message Private Information Retrieval: Capacity Results and Near-Optimal Schemes,” IEEE Transactions on Information Theory, vol. 64, no. 10, pp. 6842-6862, Oct. 2018.

[22] A. Gholami, K. Wan, H. Sun, M. Ji, G. Caire, “Coded Caching with Private Demands and Caches.” arXiv preprint arXiv:2201.11539 (2022).